

Chain rule implies Tsirelson's bound: an approach from a generalized mutual information

Eyuri Wakakuwa¹ and Mio Murao^{1,2}

¹Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

²Institute for Nano Quantum Information Electronics, The University of Tokyo, Tokyo 113-0033, Japan

E-mail: wakakuwa@eve.phys.s.u-tokyo.ac.jp

Abstract. To analyze the information theoretical derivation of Tsirelson's bound based on information causality, which is defined in terms of the efficiency of random access coding, we introduce a generalized mutual information between a classical system and a general probabilistic system. The generalization is based on the consideration of the classical information capacity of a general probabilistic system. We show that the chain rule of the generalized mutual information is essential for the derivation of Tsirelson's bound. By using the mutual information, we formulate a principle, which we call "*no-supersignalling condition*", that the assistance of nonlocal correlations does not increase the capacity of classical communication. We prove that this condition is equivalent to the no-signalling condition, and as a result, we show that information causality is essentially a matter of a single party.

1. Introduction

One of the most counterintuitive phenomena that quantum mechanics predicts is nonlocality. As it is known as the violation of the Bell inequalities [1], the statistics of the outcomes of the measurements performed at two space-like points on an entangled state can exhibit strong correlations that cannot be described within the framework of local realism. It is also known that, despite this fact, quantum correlations still satisfies the no-signalling condition, i.e., the condition that they cannot be used for superluminal communication, which is prohibited by the law of special relativity. The quantum violation of the Clauser-Horne-Shimony-Holt (CHSH) inequality [2] is limited above by Tsirelson's bound [9]. In a seminal paper [3], Popescu and Rohrlich showed that Tsirelson's bound is strictly lower than the limit imposed by the no-signalling condition alone. This result raised a question why the strength of nonlocality is limited in the quantum world. If we could find an operational principle rather than mathematical one to answer this question, it would help us better understand why quantum mechanics is the way it is [6, 7, 8].

From the information theoretical point of view, it is natural to ask if the superstrong nonlocality, i.e., the nonlocal correlations exceeding Tsirelson's bound, could enhance our ability to send classical information to a distant receiver [4]. Suppose that Alice is trying to send classical information to distant Bob by using the assistance of preshared nonlocal correlations. By the no-signalling condition, if no classical communication from Alice to Bob is performed, Bob's information gain is zero bit. In other words, zero bit of classical communication can produce no more than zero bit of classical information gain to the receiver. On the other hand, it might be possible that $m(> 0)$ bits of classical communication produces more than m bits of classical information gain to the receiver. The possibility of such an implausible situation would be related to the strength of the nonlocal correlations. Especially, one may expect that Tsirelson's bound is derived from the impossibility condition of such a situation.

Motivated by such a consideration, *information causality* is proposed as an answer to the question [4]. Information causality is the condition that “*in the bipartite nonlocality-assisted random access coding protocols, the total amount of the receiver's information gain cannot be greater than the amount of classical communication allowed in the protocol*”. This condition is never violated in the classical and the quantum theory, whereas it is violated in the theory in which the nonlocal correlation exceeding Tsirelson's bound is allowed. It implies that Tsirelson's bound is derived from this purely information theoretical principle. Thus information causality is regarded as one of the basic informational principles at the foundations of quantum mechanics.

In [4], it is proved that information causality is never violated in any no-signalling theory in so far as we can define “mutual information” satisfying five natural properties in the theory. Conversely, the violation of Tsirelson's bound, which implies the violation of information causality, indicates that at least one of the five properties of the mutual information is missed. Then it is natural to ask another question that which of the five

properties the violation of Tsirelson's bound implies.

In order to answer this question, we need to define “mutual information” in the form that is applicable to general probabilistic theories. Several investigations have been made along this line. In [18, 19], a generalized entropy H is defined in the form that is applicable to general probabilistic theories, and then a mutual information is defined in terms of the generalized entropy by $I(A : B) := H(A) + H(B) - H(A, B)$. By using this mutual information, it is proved that the data processing inequality is not satisfied in the theory in which Tsirelson's bound is violated. Similar results are also obtained in [20, 21]. However, the way of defining the entropies in their approaches are mathematical, and it is not clear whether such a definition fits into a natural extension of the concept of information in general settings. Note that in classical and quantum information theory, the operational meaning of the entropy and the mutual information is given by the source coding theorems and the channel coding theorems. In [19], a coding theorem analogous to the Schumacher compression theorem [12] in the quantum information theory is investigated using their generalized entropy. However, their consideration is only applicable under several restrictions. As it is discussed in [18], we need to seek definitions of the generalized entropies and mutual informations based on the analysis of the data compression or the channel capacities. Such an approach is also attempted in [11].

In this paper, motivated by the discussions in [18] and by the attempt in [11], we introduce an operational definition of the mutual information that is applicable to any general probabilistic theories. This is a generalization of the quantum mutual information between a classical system and a quantum system. Unlike the previous entropic approaches, we directly address to the mutual information. The generalization is based on the channel coding theorem. Thus the generalized mutual information inherently has an operational meaning as the transmission rate of classical information, or the classical information capacity of the physical system. Our definition does not require any mathematical notion such as the state space or the fine-grained measurement. The generalized mutual information is defined between a classical system and a general probabilistic system, and is not applicable to two general probabilistic systems, but it is sufficient for analyzing the situation describing information causality. The generalized mutual information satisfies four of the five properties of the mutual information except the chain rule. It automatically implies that the violation of Tsirelson's bound indicates the violation of the chain rule.

By using the generalized mutual information, we further investigate the derivation of Tsirelson's bound in terms of information causality. We formulate a principle, which we call “*no-supersignalling condition*”, that the assistance of nonlocal correlations does not increase the capacity of classical communication. We prove that this condition is equivalent to the no-signalling condition. This result is similar to the result obtained in [19], but now become operationally well corroborated. Applying this result to the analysis of information causality, we argue that information causality is not referring to the difference between the amount of classical communication and the receiver's

information gain as it was thought to be. Instead, information causality imposes an upper bound on the efficiency of random access coding, and is essentially a matter of *one* party. It turns out that the efficiency of random access coding is closely related to the chain rule. As an example for this fact, we show that we can restrict the state space of *one* gbit from the chain rule.

This paper is organized as follows. In Section 2, we give a brief review of information causality. In Section 3, we give the definition of the generalized mutual information, and show that Tsirelson's bound is derived from the chain rule. In Section 4, we prove that the generalized mutual information is indeed a generalization of the quantum mutual information. In Section 5, we give the formulation of no-supersignalling condition, and give the proof that the condition is equivalent to the no-signalling condition. In Section 6, we discuss the relation between the chain rule and random access coding, and show that information causality is a matter of one party. In Section 7, we show that we can restrict the state space of one gbit by assuming the chain rule. We conclude with discussions in Section 8.

2. Review of information causality

Information causality is the principle that “*the total amount of classical information gain that the receiver can obtain in the bipartite nonlocality-assisted random access coding protocol cannot be greater than the amount of classical communication that is allowed in the protocol*”. Suppose that a string of N random and independent bits $\vec{X} = X_1, \dots, X_n$ is given to Alice, and a random number $k \in \{1, \dots, n\}$ is given to distant Bob. The task is for Bob to correctly guess X_k under the condition that they can use a preshared resource of correlations and an m -bit one way classical communication from Alice to Bob (see Figure 1). To accomplish this task, Alice first performs a measurement on her part of the resource (denoted by A in the figure), depending on \vec{X} . She then constructs a m -bit message \vec{M} from \vec{X} and the measurement outcome, and sends it to Bob. Bob, after receiving \vec{M} , performs a measurement on his part of the resource (denoted by B in the figure), depending on \vec{M} and k . From the outcome of the measurement he computes his guess G_k for X_k . The efficiency of their protocol to accomplish this task is quantified by

$$J := \sum_{k=1}^n I_C(X_k : G_k) , \quad (1)$$

where $I_C(X_k : G_k)$ is the classical (Shannon) mutual information between X_k and G_k . Information causality is the condition that, whatever strategy they take and whatever preshared correlation allowed in the theory they use,

$$J \leq m \quad (2)$$

must hold for all $m \geq 0$. The derivation of Tsirelson's bound in terms of information causality consists of the following two theorems that are proved in [4].

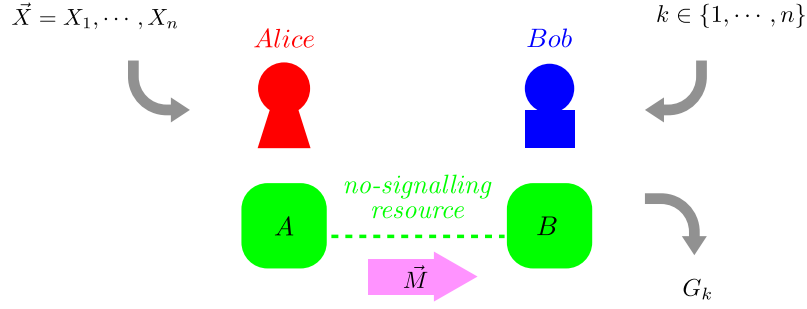


Figure 1. Nonlocality-assisted random access coding. The task is for Bob to correctly guess X_k , where k is a random number unknown to Alice.

Theorem 2.1 If we can define a function $I(A : B)$ satisfying the following properties in the general probabilistic theory, $J \leq m$ holds for all $m \geq 0$. The properties are

- *Symmetry* : $I(A : B) = I(B : A)$ for any system A and B .
- *Nonnegativity* : $I(A : B) \geq 0$ for any system A and B .
- *Consistency* : If both of the systems A and B are in classical states, $I(A : B)$ coincides with the classical mutual information.
- *Data Processing Inequality* : Under any local transformation that maps the states of system B into the state of another system B' without post-selection, $I(A : B) \geq I(A : B')$.
- *Chain Rule* : For any system A, B and C , *conditional mutual information* defined by $I(A : B|C) := I(A : B, C) - I(A : C)$ is symmetric in A and B .

Theorem 2.2 If there exists a nonlocal correlation exceeding Tsirelson's bound, we can construct a nonlocality-assisted communication protocol by which $J > m$ is achieved.

Theorem 2.1 guarantees that both of the classical theory and the quantum theory satisfy information causality. Theorem 2.2 implies that all “supernonlocal” theories, i.e., the general probabilistic theories in which the existence of nonlocal correlations stronger than Tsirelson's bound is allowed, do violate information causality. Therefore, in any supernonlocal theory, at least one of the five properties of the mutual information is missed for any definition of the mutual information. Conversely, if we could find a definition of the mutual information that is applicable for any general probabilistic theory, and if it always satisfies four of the five properties, the remaining one property could be regarded as one of the basic informational relations at the foundation of quantum mechanics.

3. Generalized mutual information

Suppose that there are a classical system X and a system S that is described by a general probabilistic theory. The states of X are labeled by a finite alphabet \mathcal{X} . For each state x of X , the corresponding state of S denoted by ϕ_x is determined.



Figure 2. The channel that we consider to define the mutual information between the system X and the system S . It has a classical system as the input system and a general probabilistic system as the output system.

The state of the composite system XS is determined by the probability distribution $p(x) = \Pr(X = x)$, which represents the probability that the system X is in the state x , and the corresponding state ϕ_x of S . Thus the state of the composite system XS is identified with an ensemble $\{p(x), \phi_x\}_{x \in \mathcal{X}}$. To define a generalized mutual information $I_G(X : S)$ between the system X and the system S in the state $\{p(x), \phi_x\}_{x \in \mathcal{X}}$, we analyze the classical information capacity of a channel that outputs the system S in the state ϕ_x according to the input $X = x$ (Figure 2). As usually considered in information theory, the sender Alice, who has access to X , tries to send classical information to the receiver Bob, who has access to S , by using the channel many times. Suppose that they use l identical and independent copies of this channel. Let X_1, \dots, X_l be the inputs of the l channels and S_1, \dots, S_l be the corresponding output systems.

Alice's encoding scheme is determined by a codebook. Let $w \in \{1, \dots, N\}$ be a message that Alice tries to communicate, and the codeword $x^l(w) = x_1(w) \cdots x_l(w)$ be the corresponding input sequence to the channels. The codebook \mathcal{C} is defined as the list of the codewords for all messages by

$$\mathcal{C} := \begin{bmatrix} x_1(1) & \cdots & x_l(1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_l(N) \end{bmatrix}. \quad (3)$$

The letter frequency $f(x)$ for the codebook is defined by

$$f(x) := \frac{|\{(k, w) | x_k(w) = x, 1 \leq k \leq l, 1 \leq w \leq N\}|}{lN} \quad (x \in \mathcal{X}). \quad (4)$$

For the given probability distribution $\{p(x)\}_{x \in \mathcal{X}}$, the tolerance τ of the code is defined by

$$\tau := \max_{x \in \mathcal{X}} |p(x) - f(x)|. \quad (5)$$

By performing a decoding measurement on the output systems S_1, \dots, S_l , Bob tries to guess what the original message w is. Let \mathcal{D} denote the decoding measurement. Note that, in general, the decoding measurement is not one in which Bob performs a measurement on each of S_1, \dots, S_l individually, but one in which the whole of the composite system $S_1 \cdots S_l$ is subjected to a measurement. Let W, \hat{W} be Alice's original message and Bob's decoding outcome, respectively. The average error probability P_e is defined by

$$P_e := \frac{1}{N} \sum_{u=1}^N \Pr(\hat{W} \neq u | W = u). \quad (6)$$

The pair of the codebook \mathcal{C} and the decoding measurement \mathcal{D} is called an (N, l) code. The ratio $\log N/l$ is called the rate of the code, and it represents how many bits of classical information is transmitted per use of the channel.

Definition 3.1 A rate R is said to be achievable with $p(x)$ if there exists a sequence of $(2^{lR}, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ such that

- (i) $P_e^{(l)} \rightarrow 0$ when $l \rightarrow \infty$,
- (ii) $\tau^{(l)} \rightarrow 0$ when $l \rightarrow \infty$,
- (iii) All codeletters $x_k(w)$ in $\mathcal{C}^{(l)}$ are elements of $\bar{\mathcal{X}} := \text{supp } p = \{x | x \in \mathcal{X}, p(x) \neq 0\}$.

Definition 3.2 The mutual information between a classical system X and a general probabilistic system S , denoted by $I_G(X : S)$, is the function that satisfies the condition that

- (i) A rate R is achievable with $p(x)$ if $R < I_G(X : S)$,
- (ii) A rate R is achievable with $p(x)$ only if $R \leq I_G(X : S)$.

We also define $I_G(S : X)$ by $I_G(S : X) := I_G(X : S)$.

Theorem 3.3 $I_G(X : S)$ exists and satisfies $I_G(X : S) \leq H(X)$. Here, $H(X)$ is the Shannon entropy of the system X defined by $H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$.

Proof. First we prove the existence of $R^* := \sup \{R | R \text{ is achievable with } p(x)\}$. Consider a $(2^{lR}, l)$ code and suppose that Alice's message $W = 1, \dots, 2^{lR}$ is uniformly distributed. Let I', H' be the mutual information and the entropy when the input sequence is the codewords corresponding to the uniformly distributed message W . By Fano's inequality, we have

$$H'(W|\hat{W}) \leq P_e^{(l)} lR + 1 \quad (7)$$

where $P_e^{(l)} = P(W \neq \hat{W})$. Thus

$$\begin{aligned} lR = H'(W) &= I'(W : \hat{W}) + H'(W|\hat{W}) \\ &\leq I'(X^l : \hat{W}) + P_e^{(l)} lR + 1 \\ &\leq H'(X^l) + P_e^{(l)} lR + 1. \end{aligned} \quad (8)$$

Here, we use the data processing inequality in the first inequality. By introducing a classical variable K that indicates k with the probability distribution $P(K = k) = 1/l$, we also have

$$H'(X^l) \leq \sum_{k=1}^l H'(X_k) = lH'(X|K) \leq lH'(X). \quad (9)$$

From (8) and (9), we obtain

$$P_e^{(l)} \geq 1 - \frac{H'(X)}{R} - \frac{1}{lR}. \quad (10)$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes satisfying $P_e^{(l)} \rightarrow 0$ and $H'(X) \rightarrow H(X)$ when $l \rightarrow \infty$. Thus $R \leq H(X)$. Hence R^* exists and satisfies $R^* \leq H(X)$.

Next we prove that any rate $R < R^*$ is also achievable with $p(x)$. Let $\{(\mathcal{C}^{*(l)}, \mathcal{D}^{*(l)})\}_l$ be the sequence of $(2^{lR^*}, l)$ codes with all codeletters in $\bar{\mathcal{X}}$ that satisfies $P_e^{*(l)} \rightarrow 0$ and $\tau^{*(l)} \rightarrow 0$. For arbitrary $0 \leq \lambda < 1$, define another codebook $\mathcal{C}^{(l)}$ by using $\mathcal{C}^{*(\lambda l)}$ for the first λl codeletters and by choosing the last $(1 - \lambda)l$ codeletters arbitrarily so that the total tolerance is sufficiently small. Also define the corresponding decoding measurement $\mathcal{D}^{(l)}$ as the measurement in which the output system $S_1 \cdots S_{\lambda l}$ is subjected to the decoding measurement $\mathcal{D}^{*(l)}$ and the output system $S_{\lambda l+1} \cdots S_l$ is ignored. The code sequence $\{(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})\}_l$ constructed in this way is a sequence of $(2^{l\lambda R^*}, l)$ codes with all codeletters in $\bar{\mathcal{X}}$ that satisfies $P_e^{(l)} \rightarrow 0$ and $\tau^{(l)} \rightarrow 0$. Thus $R = \lambda R^*$ is achievable with $p(x)$. Hence we obtain $R^* = I_G(X : S)$. \square

Note that $I_G(X : S)$ is a function of the state $\Gamma := \{p(x), \phi_x\}_{x \in \mathcal{X}}$ of the composite system XS . To emphasize this, we sometimes use the notation $I_G(X : S)_\Gamma$. Since $R = 0$ is always achievable, $I_G(X : S)$ is nonnegative. Shannon's noisy channel coding theorem guarantees that $I_G(X : S)$ coincides with the classical mutual information $I_C(X : S)$ if S is a classical system. The generalized mutual information satisfies the data processing inequality as follows.

Property 3.4 Let $\mathcal{E}_{S \rightarrow S'}$ be any local transformation that maps the states of a general probabilistic system S into the state of another general probabilistic system S' . If $\mathcal{E}_{S \rightarrow S'}$ contains no post-selection, the generalized mutual information does not increase under this transformation, i.e., $I_G(X : S) \geq I_G(X : S')$. Similarly, $I_G(X : S) \geq I_G(X' : S)$ under any local transformation $\mathcal{E}_{X \rightarrow X'}$ that maps the states of a classical system X into the state of another classical system X' without post-selection.

Proof. Here we only prove the former part. For the latter part, see Appendix A. Consider two channels, the channel I and the channel II (see Figure 3). Depending on the input $X = x$, the channel I emits the system S in the state ϕ_x , and the channel II emits the system S' in the state $\phi'_x = \mathcal{E}_{S \rightarrow S'}(\phi_x)$. It is only necessary to verify that if a rate R is achievable with $p(x)$ by the channel II, R is also achievable with $p(x)$ by the channel I. Let $\{(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})\}_l$ be a sequence of $(2^{lR}, l)$ codes for the channel II with the average error probability $P_e'^{(l)}$ and the tolerance $\tau'^{(l)}$. From the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$, construct a $(2^{lR}, l)$ code $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for the channel I by $\mathcal{C}^{(l)} = \mathcal{C}'^{(l)}$ and $\mathcal{D}^{(l)} = \mathcal{D}'^{(l)} \circ \mathcal{E}_{S \rightarrow S'}^{\otimes l}$. Here, $\mathcal{D}'^{(l)} \circ \mathcal{E}_{S \rightarrow S'}^{\otimes l}$ represents a process in which first $\mathcal{E}_{S \rightarrow S'}$ is applied to each of S_1, \dots, S_l individually and then the decoding measurement $\mathcal{D}'^{(l)}$ is performed on the total output system $S'_1 \cdots S'_l$. The average error probability and the tolerance of this code are given by $P_e^{(l)} = P_e'^{(l)}$ and $\tau^{(l)} = \tau'^{(l)}$, respectively. Hence, if $P_e'^{(l)} \rightarrow 0$ and $\tau'^{(l)} \rightarrow 0$, we also have $P_e^{(l)} \rightarrow 0$ and $\tau^{(l)} \rightarrow 0$, and thus R is achievable with $p(x)$ by the channel I. \square

In the general probabilistic theories, a measurement on a system S without post-selection is described by a probabilistic map \mathcal{E}_M that maps the states of S into the states

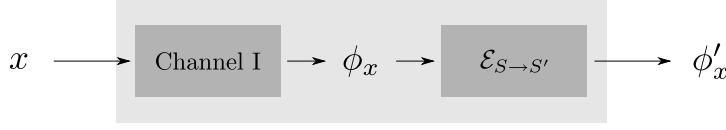


Figure 3. The channel II defined as the combination of the channel I and $\mathcal{E}_{S \rightarrow S'}$.

of a classical system T_S . T_S represents the register of the measurement outcomes. As a special case for Property 3.4, we have $I_G(X : T_S) \leq I_G(X : S)$ under \mathcal{E}_M , which is a generalization of Holevo's inequality. Let us define the accessible information $I_{acc}(X : S)$ by

$$I_{acc}(X : S) := \max I_C(X : T_S), \quad (11)$$

where the maximization is taken over all possible measurements on S . Then we have $0 \leq I_{acc}(X : S) \leq I_G(X : S)$.

Property 3.5 $I_{acc}(X : S) = 0$ if and only if $I_G(X : S) = 0$.

Proof. The “if” part is a direct consequence of Property 3.4. To prove the “only if” part, note that $I_{acc}(X : S) = 0$ implies that ϕ_x is the same state for all $x \in \bar{\mathcal{X}}$. Thus all states $\phi_{\mathbf{x}} := \phi_{x_1} \phi_{x_2} \cdots \phi_{x_l}$ ($\mathbf{x} \in \bar{\mathcal{X}}^l$) are the same states of the composite system $S_1 S_2 \cdots S_l$. It implies that no code with all codeletters in $\bar{\mathcal{X}}$ can be used for transmitting information. \square

To summarize, our generalized mutual information satisfies the following properties.

- *Symmetry:* $I_G(X : S) = I_G(S : X)$.
- *Nonnegativity:* $I_G(X : S) \geq 0$
- *Consistency:* When S is a classical system, $I_G(X : S) = I_C(X : S)$.
- *Data Processing Inequality:* $I_G(X : S) \geq I_G(X' : S')$ under local stochastic maps $\mathcal{E}_{X \rightarrow X'}$ and $\mathcal{E}_{S \rightarrow S'}$ that contain no post-selection.

Thus, from Theorem 2.1 and Theorem 2.2, we conclude that the chain rule of the generalized mutual information should be violated in any supernonlocal theory. Conversely, the chain rule implies Tsirelson's bound.

4. Quantum mutual information

The generalized mutual information defined by Definition 3.2 looks different from the quantum mutual information $I_Q(X : S)$ defined by

$$I_Q(X : S)_\rho = H(S)_{\bar{\rho}} - \sum_{x \in \mathcal{X}} p(x) H(S)_{\rho_x}, \quad (12)$$

where

$$\rho = \sum_{x \in \mathcal{X}} p(x) |x\rangle\langle x|^X \otimes \rho_x^S \quad (13)$$

and

$$\bar{\rho} = \sum_{x \in \mathcal{X}} p(x) \rho_x . \quad (14)$$

However, with a slight generalization of the Holevo-Schumacher-Westmoreland theorem, it is shown that these definitions are equivalent in quantum mechanics.

Theorem 4.1 In quantum mechanics, the generalized mutual information coincides with the quantum mutual information, i.e.,

$$I_G(X : S)_{\Gamma_\rho} = I_Q(X : S)_\rho \quad (15)$$

where

$$\rho = \sum_{x \in \mathcal{X}} p(x) |x\rangle\langle x|^X \otimes \rho_x^S \quad (16)$$

and $\Gamma_\rho = \{p(x), \rho_x\}_{x \in \mathcal{X}}$.

Proof. To prove this, it is only necessary to verify the following two statements:

- (i) A rate R is achievable with $p(x)$ if $R < I_Q(X : S)_\rho$,
- (ii) A rate R is achievable with $p(x)$ only if $R \leq I_Q(X : S)_\rho$.

The first statement is proved in [13, 14] by using random code generation, and the second statement is proved in the following way. Consider a $(2^{lR}, l)$ code and suppose that Alice's message $W = 1, \dots, 2^{lR}$ is uniformly distributed. Similarly to (8), we have

$$lR = H'(W) = I'(W : \hat{W}) + H'(W | \hat{W}) \leq I'_Q(X^l : S^l) + P_e^{(l)} lR + 1 . \quad (17)$$

Here, we use the data processing inequality. We also have

$$\begin{aligned} I'_Q(X^l : S^l) &= H'(S^l) - H'(S^l | X^l) = H'(S^l) - \sum_{k=1}^l H'(S_k | X_k) \\ &\leq \sum_{k=1}^l (H'(S_k) - H'(S_k | X_k)) = \sum_{k=1}^l I'_Q(X_k : S_k) \\ &= lI'_Q(X : S | K) = lI'_Q(X, K : S) - lI'_Q(K : S) \\ &\leq lI'_Q(X, K : S) = lI'_Q(X : S) . \end{aligned} \quad (18)$$

In the first line, we use the fact that the state of S_k depends only on X_k . The first inequality is from the subadditivity of the von Neumann entropy. The last equality holds since $K \rightarrow X \rightarrow S$ forms a Markov chain. From (17) and (18), we obtain

$$P_e^{(l)} \geq 1 - \frac{I'_Q(X : S)}{R} - \frac{1}{lR} . \quad (19)$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes satisfying $P_e^{(l)} \rightarrow 0$ and $I'_Q(X : S) \rightarrow I_Q(X : S)_\rho$ when $l \rightarrow \infty$. Thus $R \leq I_Q(X : S)_\rho$. \square

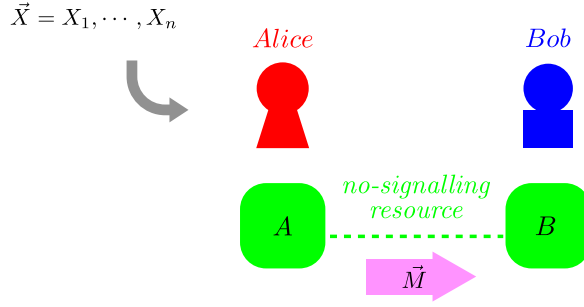


Figure 4. The situation that the no-supersignalling condition refers to. The amount of information about \vec{X} contained in \vec{M} and B is quantified by $I_G(\vec{X}, \vec{M}, B)$.

5. No-supersignalling condition

In this section, we formulate a principle that we call “no-supersignalling condition” to further investigate the argument of information causality. Suppose that Alice is trying to send to distant Bob information about n independent classical bits X_1, \dots, X_n , under the condition that they can only use an m bit classical communication \vec{M} from Alice to Bob and a supplementary resource of correlations preshared between them (see Figure 4). The situation is similar to the setting of information causality described in Section 2, but now, we do not introduce random access coding. Instead, we evaluate Bob’s information gain by $I_G(\vec{x} : \vec{M}, B)$. We define that no-supersignalling condition is satisfied if $I_G(\vec{X} : \vec{M}, B) \leq m$ holds for all $m \geq 0$. The condition indicates that “*the assistance of correlations cannot increase the capacity of classical communication*”, which was the *original* motivation for introducing information causality. In what follows, we prove that the no-supersignalling condition is equivalent to the no-signalling condition.

Lemma 5.1 For any classical system X, Y and any general probabilistic system S , if $I_{acc}(X : S) = 0$ then $I_{acc}(X : S, Y) \leq H(Y)$.

Proof. Consider a channel with the input system X and the output system S, Y (see Figure 5). Let \mathcal{Z} be the set of all measurements on S , and $p(t|x, y, z)$ be the probability of obtaining the outcome t when the measurement $z \in \mathcal{Z}$ is performed on the system S in the state ϕ_{xy} . To achieve $I_{acc}(X : S, Y)$, the receiver performs a measurement on S possibly depending on Y . Let $z(y)$ be the optimal choice of the measurement when $Y = y$. The probability of obtaining the outcome t when $X = x$ and $Y = y$ is given by

$$p_1(t|x, y) := p(t|x, y, z(y)) . \quad (20)$$

We define

$$p_1(t, x, y) := p(x, y)p_1(t|x, y) = p(x, y)p(t|x, y, z(y)) . \quad (21)$$

The condition $I_{acc}(X : S) = 0$ implies that for all $z \in \mathcal{Z}$,

$$\sum_y p(x, y)p(t|x, y, z) = p(x)p_2(t|z) , \quad (22)$$



Figure 5. The channel that we consider to prove Lemma 5.1. For each pair of the input $X = x$ and the output $Y = y$, the corresponding state ϕ_{xy} of the output system S is determined.

where

$$p_2(t|z) := \sum_{x,y} p(x,y)p(t|x,y,z) . \quad (23)$$

Thus we obtain

$$\begin{aligned} p_1(t, x, y) &= p(x, y)p(t|x, y, z(y)) \\ &\leq \sum_{y'} p(x, y')p(t|x, y', z(y)) \\ &= p(x)p_2(t|z(y)) . \end{aligned} \quad (24)$$

The accessible information $I_{acc}(X : S, Y)$ is equal to the mutual information $I_C(X : T, Y)$ calculated for the probability distribution $p_1(t, x, y)$. Therefore

$$\begin{aligned} I_{acc}(X : S, Y) &= I_C(X : T, Y)_{p_1} \\ &= \sum_{t,x,y} p_1(t, x, y) \log \frac{p_1(t, x, y)}{p(x)p_1(t, y)} \\ &= H(Y) + \sum_{t,x,y} p_1(t, x, y) \log \frac{p_1(t, x, y)p(y)}{p(x)p_1(t, y)} \\ &\leq H(Y) + \sum_{t,x,y} p_1(t, x, y) \log \frac{p(x)p(y)p_2(t|z(y))}{p(x)p_1(t, y)} \\ &= H(Y) - \sum_{t,y} p_1(t, y) \log \frac{p_1(t, y)}{p_2(t, y)} \\ &= H(Y) - D(p_1(t, y) \| p_2(t, y)) \\ &\leq H(Y) . \end{aligned}$$

In the first inequality, we used (24). In the next equality we defined a probability distribution $p_2(t, y) := p_2(t|z(y))p(y)$. The last inequality is from the nonnegativity of the relative entropy. \square

Theorem 5.2 For any classical system X , Y and any general probabilistic system S , if $I_G(X : S) = 0$ then $I_G(X : S, Y) \leq H(Y)$.

Proof. Consider a $(2^{lR}, l)$ code with all codeletters in $\bar{\mathcal{X}}$ for the channel shown in Figure 5, and suppose that Alice's message is uniformly distributed. By Fano's inequality, we have

$$I'(W : \hat{W}) \geq lR - 1 - P_e^{(l)}lR . \quad (25)$$

By the data processing inequality, we also have

$$I'(W : \hat{W}) \leq I'(X^l : Y^l, T_{S^l}) \leq I'_{acc}(X^l : Y^l, S^l). \quad (26)$$

From Property 3.5, the condition $I_G(X : S) = 0$ implies $I'_{acc}(X^l : S^l) = 0$. From Lemma 5.1, we obtain

$$I'_{acc}(X^l : Y^l, S^l) \leq H'(Y^l), \quad (27)$$

and thus

$$I'(W : \hat{W}) \leq H'(Y^l) \leq lH'(Y). \quad (28)$$

Hence we obtain

$$(1 - P_e^{(l)})R \leq H'(Y) + \frac{1}{l}. \quad (29)$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes with all codeletters in $\bar{\mathcal{X}}$ that satisfies $P_e^{(l)} \rightarrow 0$ and $H'(Y) \rightarrow H(Y)$ when $l \rightarrow \infty$. Thus, for any R that is achievable with $p(x)$, we have $R \leq H(Y)$. It implies $I_G(X : Y, S) \leq H(Y)$. \square

Corollary 5.3 No-supersignalling condition is equivalent to the no-signalling condition.

Proof. By setting $X = \vec{X}$ and $Y = \vec{M}$ in the result of Theorem 5.2, for all $m \geq 0$, we obtain $I_G(\vec{X} : \vec{M}, B) \leq H(\vec{M}) \leq m$ from the no-signalling condition. Note that $I_{acc}(\vec{X} : B) = 0$ is required by the no-signalling condition and thus $I_G(\vec{X} : B) = 0$ by Property 3.5. Conversely, for $m = 0$, no-supersignalling condition that $I_G(X : B) = 0$, and it is equal to the no-signalling condition. \square

6. The chain rule and random access coding

In this section, by using the result obtained in Section 5, we discuss the relation among information causality, random access coding and the chain rule. Let us define

$$\Delta_{\text{NSS}} := I_G(\vec{X} : \vec{M}, B) - m, \quad (30)$$

$$\Delta_{\text{RAC}} := J - I_G(\vec{X} : \vec{M}, B), \quad (31)$$

$$\Delta_{\text{IC}} := \Delta_{\text{NSS}} + \Delta_{\text{RAC}} = J - m. \quad (32)$$

Δ_{NSS} quantifies the capacity of classical communication assisted by correlations, and Δ_{RAC} quantifies the efficiency of random access coding. No-supersignalling condition is equivalent to $\Delta_{\text{NSS}} \leq 0$ and information causality is equivalent to $\Delta_{\text{IC}} \leq 0$.

Theorem 2.2 states that, if Tsirelson's bound is violated, we have $\Delta_{\text{IC}} > 0$. Therefore the violation of Tsirelson's bound implies at least either $\Delta_{\text{NSS}} > 0$ or $\Delta_{\text{RAC}} > 0$. Then we would ask the following question: which does the violation of Tsirelson's bound imply, $\Delta_{\text{NSS}} > 0$ or $\Delta_{\text{RAC}} > 0$? As we proved in Section 5, $\Delta_{\text{NSS}} \leq 0$ is satisfied by all no-signalling theories. Thus the answer is that the violation of Tsirelson's bound only implies $\Delta_{\text{RAC}} > 0$. Therefore, information causality does not refer to the difference between the amount of classical communication and the receiver's information

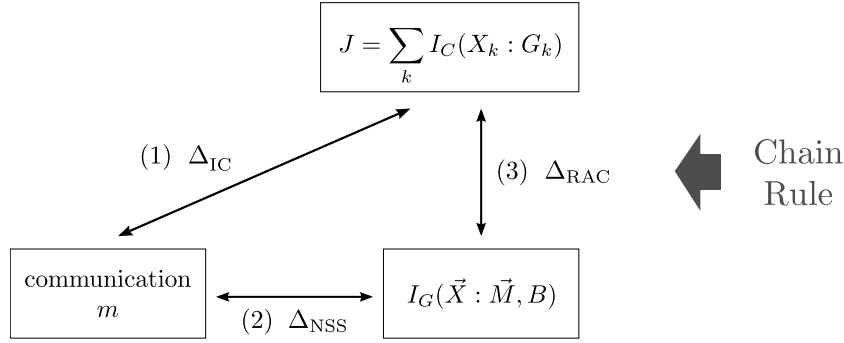


Figure 6. The relation among information causality, the no-supersignalling or the no-signalling condition and random access coding. Information causality refers to the gap in (1) represented by Δ_{IC} . No-supersignalling refers to the gap in (2) represented by Δ_{IC} , and it is irrelevant to Tsirelson's bound. The gap in (3) represented by Δ_{RAC} is crucial for the derivation of Tsirelson's bound. Δ_{RAC} is bounded above due to the chain rule of the generalized mutual information.

gain. Information causality refers to how much random access coding can be efficient, and is essentially a matter of *one* party. In the derivation of Tsirelson's bound, non-existence of super-efficient RAC represented by $\Delta_{RAC} \leq 0$ is critical [16] (see Figure 6).

It is proved in [4] that, if the assumption in Theorem 2.1 is satisfied, we have $\Delta_{RAC} \leq 0$. Since our generalized mutual information satisfies all those properties except the chain rule, $\Delta_{RAC} > 0$ implies the violation of the chain rule. Let X, Y be two classical systems and S be a general probabilistic system. The chain rule of the generalized mutual information is given by

$$I_G(X, Y : S) + I_G(X : Y) = I_G(X : S) + I_G(Y : S, X) . \quad (33)$$

Especially, in the case where $I_G(X : Y) = 0$, we have

$$I_G(X, Y : S) = I_G(X : S) + I_G(Y : S, X) . \quad (34)$$

Each term in (34) has an operational meaning as the information transmission rate by definition. The relation is satisfied in both classical and quantum theory, but is violated in all supernonlocal theories. Thus we can conclude that this highly nontrivial relation gives a strong restriction on the physical theories. However, the operational meaning of this relation is not clear so far.

7. Restriction on one gbit state space

To investigate the relationship between random access coding and the chain rule, we consider a gbit. A gbit is the counterpart of a qubit in general probabilistic theories [17]. Here, we assume no property for a gbit such as the dimension of the state space, the possibility or impossibility of various measurements and transformations. Instead, we define a gbit as the minimum unit of information in the theory, and we require that

the classical information capacity of one gbit is not more than one bit. Thus we require

$$I_G(X : S_{\text{1gb}}) \leq 1 \quad (35)$$

for any classical system X . When X is a classical system composed of two independent and uniformly random bits X_0 and X_1 , we have

$$I_G(X_0, X_1 : S_{\text{1gb}}) \leq 1. \quad (36)$$

By the chain rule, we have

$$I_G(X_0, X_1 : S_{\text{1gb}}) = I_G(X_0 : S_{\text{1gb}}) + I_G(X_1 : S_{\text{1gb}}, X_0). \quad (37)$$

By the data processing inequality, we also have

$$I_G(X_0 : S_{\text{1gb}}) + I_G(X_1 : S_{\text{1gb}}, X_0) \geq I_{\text{acc}}(X_0 : S_{\text{1gb}}) + I_{\text{acc}}(X_1 : S_{\text{1gb}}). \quad (38)$$

Thus the chain rule implies

$$I_{\text{acc}}(X_0 : S_{\text{1gb}}) + I_{\text{acc}}(X_1 : S_{\text{1gb}}) \leq 1. \quad (39)$$

We consider the success probabilities of the decoding measurements on S_{1gb} for X_0 and X_1 . For simplicity, we assume that the optimal measurement performed on S_{1gb} to decode X_0 or X_1 has two outcomes $t = 0, 1$. Let $P(t|m, x_0, x_1)$ be the probability of obtaining the outcome t when $X_0 = x_0$, $X_1 = x_1$ and the measurement m is performed. The index $m = 0, 1$ corresponds to the optimal measurement for decoding X_0 , X_1 , respectively. The list of all probabilities $\{P(t|m, x_0, x_1)\}_{t,m,x_0,x_1=0,1}$ can be regarded as representing a “state”. We compare the state space of a qubit and the state space determined by (39). For further simplicity, we assume that for all x_0 and x_1 ,

$$\begin{aligned} P(t = x_0|m = 0, x_0, x_1) &= \frac{1 + \alpha}{2} \quad (0 \leq \alpha \leq 1), \\ P(t = x_1|m = 1, x_0, x_1) &= \frac{1 + \beta}{2} \quad (0 \leq \beta \leq 1). \end{aligned}$$

Then we have

$$\begin{aligned} I_{\text{acc}}(X_0 : S_{\text{1gb}}) &= I_C(x_1 : t|m = 0) = 1 - H(x_0|t, m = 0) \\ &= 1 - H(x_0 \oplus t|m = 0) = 1 - h\left(\frac{1 + \alpha}{2}\right), \end{aligned} \quad (40)$$

and

$$I_{\text{acc}}(X_1 : S_{\text{1gb}}) = 1 - h\left(\frac{1 + \beta}{2}\right). \quad (41)$$

Here, $h(x)$ is the binary entropy defined by $h(x) = -x \log x - (1 - x) \log (1 - x)$. From (39), (40) and (41), we have

$$h\left(\frac{1 + \alpha}{2}\right) + h\left(\frac{1 + \beta}{2}\right) \geq 1. \quad (42)$$

This inequality gives a nontrivial restriction on the state space of one gbit (see Figure 7). It implies that the chain rule imposes a restriction on the possibility of “superstrong” random access coding. It is shown in Appendix B that in the case of one qubit, the obtainable region is given by $\alpha^2 + \beta^2 \leq 1$.

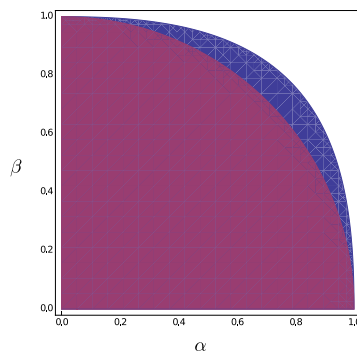


Figure 7. Comparison of the state space of a qubit and the boundary given by the chain rule. The red region indicates the state space of a qubit given by $\alpha^2 + \beta^2 \leq 1$. The blue region in addition to the red region indicates the region defined by (42).

8. Conclusions and discussions

We have defined a generalized mutual information between a classical system and a general probabilistic system. Since the definition is based on the channel coding theorem, our generalized mutual information inherently has an operational meaning as the information transmission rate. We have shown that the mutual information coincides with the quantum mutual information if the system is quantum. The generalized mutual information satisfies nonnegativity, symmetry, the data processing inequality, and the consistency with the classical mutual information. However, it does not always satisfy the chain rule.

By using the generalized mutual information, we have analyzed the derivation of Tsirelson's bound based on information causality defined in terms of the efficiency of random access coding. We showed that the chain rule of the mutual information, which is satisfied by both classical and quantum theory, is violated in any theory in which the existence of nonlocal correlations exceeding Tsirelson's bound is allowed. Thus we conclude that the chain rule implies Tsirelson's bound.

We formulated a condition (no-supersignalling condition) that the assistance of preshared correlation cannot increase the capacity of the classical communication. We proved that this condition is equivalent to the no-signalling condition. Based on this result, we argued that information causality is essentially a matter of *one* party referring to the efficiency of random access coding. The efficiency of random access coding is restricted by the chain rule of the mutual information. As an example for this fact, we derived a restriction on the state space of *one* gbit from the chain rule.

Although the operational meaning of the generalized mutual information is clear, we have not yet succeeded in finding out a clear operational meaning of the chain rule. In classical and quantum Shannon theory, the chain rule appears in a lot of proofs of coding theorems. Our result shows that it is a highly nontrivial fact that the chain rule is satisfied in classical and quantum theory. Therefore, investigation of the meaning of the chain rule would lead us to a more profound understanding of the informational

foundations of quantum mechanics.

On the other hand, our definition of the generalized mutual information would not be the only way to generalize the quantum mutual information. It would also be fruitful to seek for other operationally motivated definitions of the generalized mutual information and compare them with each other.

Acknowledgments

We thank Salman Beigi and Takanori Sugiyama for useful discussions. This work was supported by Project for Developing Innovation Systems of Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. MM acknowledges support from JSPS by KAKENHI (Grant No. 23540463).

Appendix A. Data processing inequality

We prove the latter part of Theorem 3.4, which states that under any local stochastic map $\mathcal{E}_{X \rightarrow X'}$ that contains no post-selection, we have

$$I_G(X : S) \geq I_G(X' : S). \quad (\text{A.1})$$

The effect of $\mathcal{E}_{X \rightarrow X'}$ is determined by the conditional probability distribution $p_{\mathcal{E}}(x'|x)$, where x and x' denote the states of X and X' , respectively. Let $\{p(x), \phi_x\}_{x \in \mathcal{X}}$ be the state of XS before applying $\mathcal{E}_{X \rightarrow X'}$. We can define probability distributions $p_{\mathcal{E}}(x, x') = p(x)p_{\mathcal{E}}(x'|x)$, $p(x') = \sum_x p_{\mathcal{E}}(x, x')$ and $p_{\mathcal{E}}(x|x') = p_{\mathcal{E}}(x, x')/p(x')$ for $x \in \mathcal{X}$ and $x' \in \bar{\mathcal{X}}' = \{x'|x' \in \mathcal{X}', p(x') \neq 0\}$. Note that $p_{\mathcal{E}}(x|x') = 0$ for $x \notin \bar{\mathcal{X}} = \{x|x \in \mathcal{X}, p(x) \neq 0\}$. The state of $X'S$ after applying $\mathcal{E}_{X \rightarrow X'}$ is $\{p(x'), \phi_{x'}\}_{x' \in \bar{\mathcal{X}}'}$, where $\phi_{x'}$ is the mixture of ϕ_x with probability given by $p_{\mathcal{E}}(x|x')$. We assume that $|\mathcal{X}|, |\mathcal{X}'| < \infty$.

To prove (A.1), consider two channels, the channel I and the channel III (see Figure A1). The channel I outputs the system S in the state ϕ_x according to the input $X = x$, and the channel III outputs the system S in the state $\phi_{x'}$ according to the input $X' = x'$. It is only necessary to show that if a rate R is achievable with $p(x')$ by the channel III, R is also achievable with $p(x)$ by the channel I. Consider a sequence of $(2^{lR}, l)$ codes $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$ for the channel III that satisfies

- (i) $P_e'^{(l)} \rightarrow 0$ when $l \rightarrow \infty$,
- (ii) $\tau'^{(l)} \rightarrow 0$ when $l \rightarrow \infty$,
- (iii) All codeletters in $\mathcal{C}'^{(l)}$ are elements of $\bar{\mathcal{X}}'$.

Such a sequence exists if R is achievable with $p(x')$ by the channel III. From the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$, we randomly construct $(2^{lR}, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for the channel I in the following way.

- For any w and k ($1 \leq w \leq 2^{lR}, 1 \leq k \leq l$), generate the codeletter $x_k(w)$ randomly and independently according to the probability distribution $P(x_k(w) = x) = p_{\mathcal{E}}(x|x'_k(w))$.

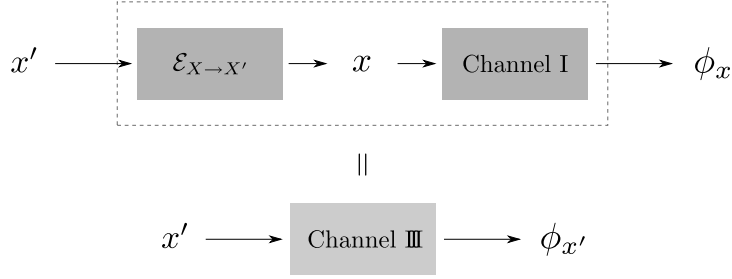


Figure A1. The channel III defined as the combination of $\mathcal{E}_{X \rightarrow X'}$ and the channel I. This channel as a whole is equivalent to a channel with the input x' and the output $\phi_{x'}$.

- Regardless of the randomly generated codebook $\mathcal{C}^{(l)}$, use the same decoding measurement $\mathcal{D}^{(l)} = \mathcal{D}'^{(l)}$.

Let $P_e^{\mathcal{C}^{(l)}}$ be the average error probability of the code $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ defined by

$$P_e^{\mathcal{C}^{(l)}} := \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) . \quad (\text{A.2})$$

Averaging $P_e^{\mathcal{C}^{(l)}}$ over all codebooks $\mathcal{C}^{(l)}$ that are randomly generated, we obtain

$$\bar{P}_e^{(l)} := \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P_e^{\mathcal{C}^{(l)}} , \quad (\text{A.3})$$

where $P(\mathcal{C}^{(l)})$ is the probability of obtaining the codebook $\mathcal{C}^{(l)}$ as a result of random code generation. In Lemma A.1, we show that $\bar{P}_e^{(l)} \rightarrow 0$ in the limit of $l \rightarrow \infty$. In Lemma A.2, we prove that for sufficiently large l , the tolerance $\tau^{(l)}$ of the codebook $\mathcal{C}^{(l)}$ is almost equal to 0 with arbitrarily high probability. Finally, we give the proof for (A.1) in Theorem A.3.

Lemma A.1

$$\lim_{l \rightarrow \infty} \bar{P}_e^{(l)} = 0 . \quad (\text{A.4})$$

Proof. $\bar{P}_e^{(l)}$ defined by (A.3) is calculated to

$$\begin{aligned} \bar{P}_e^{(l)} &= \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) \times \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) \\ &= \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) \\ &= \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} \bar{P}(\hat{W} \neq u | W = u) , \end{aligned} \quad (\text{A.5})$$

where

$$\bar{P}(\hat{W} \neq u | W = u) := \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) . \quad (\text{A.6})$$

The codebook $\mathcal{C}^{(l)}$ is determined by the codeletters $x_k(w)$ ($1 \leq w \leq 2^{lR}, 1 \leq k \leq l$). Due to the way of randomly generating the code, the probability of obtaining the codebook $\mathcal{C}^{(l)}$ such that $x_k(w) = \xi_{wk}$ ($1 \leq w \leq 2^{lR}, 1 \leq k \leq l$) is given by

$$\begin{aligned} P(\mathcal{C}^{(l)}) &= P(\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= \prod_{w=1}^{2^{lR}} \prod_{k=1}^l P(x_k(w) = \xi_{wk}) \\ &= \prod_{w=1}^{2^{lR}} \prod_{k=1}^l p_{\mathcal{E}}(x = \xi_{wk} | x' = x'_k(w)) . \end{aligned} \quad (\text{A.7})$$

Let $D(\phi_{x_1} \cdots \phi_{x_l})$ be the result of the decoding measurement $\mathcal{D}^{(l)}$ on the composite system $S_1 \cdots S_l$ in the state $\phi_{x_1} \cdots \phi_{x_l}$. We have

$$\begin{aligned} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) &= P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u | \{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u) , \end{aligned} \quad (\text{A.8})$$

and we obtain

$$\begin{aligned} &\bar{P}(\hat{W} \neq u | W = u) \\ &= \sum_{\{\xi_{wk}\}_{w,k}} P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u | \{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \times P(\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= \sum_{\{\xi_{wk}\}_k} P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u) \times P(\{x_k(u)\}_k = \{\xi_{wk}\}_k) \\ &= \sum_{\{\xi_{wk}\}_k} P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u) \times \prod_{k=1}^l p_{\mathcal{E}}(x = \xi_{wk} | x' = x'_k(u)) . \end{aligned} \quad (\text{A.9})$$

On the other hand, the error probability for the message w when the channel III is used with the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$ is given by

$$\begin{aligned} &P'(\hat{W} \neq u | W = u) \\ &= P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u) \\ &= \sum_{\{x_k\}_k} \prod_{k=1}^l p_{\mathcal{E}}(x = x_k | x' = x'_k(w)) \times P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u) . \end{aligned} \quad (\text{A.10})$$

From (A.9) and (A.10), we obtain

$$\bar{P}(\hat{W} \neq u | W = u) = P'(\hat{W} \neq u | W = u) , \quad (\text{A.11})$$

and consequently

$$\bar{P}_e^{(l)} = P_e'^{(l)} . \quad (\text{A.12})$$

Therefore $\bar{P}_e^{(l)} \rightarrow 0$ when $l \rightarrow \infty$. \square

Lemma A.2 $\tau^{(l)} \rightarrow 0$ in probability in the limit of $l \rightarrow \infty$.

Proof. Let $f(x)^{(l)}$ and $f(x')^{(l)}$ be the letter frequency of the codebook $\mathcal{C}^{(l)}$ and $\mathcal{C}'^{(l)}$, respectively. We have

$$\begin{aligned} |f(x)^{(l)} - p(x)| &= \left| f(x)^{(l)} - \sum_{x' \in \bar{\mathcal{X}}'} p_{\mathcal{E}}(x|x') p(x') \right| \\ &\leq \left| f(x)^{(l)} - \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} p_{\mathcal{E}}(x|x') \right| + \left| \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} p_{\mathcal{E}}(x|x') - \sum_{x' \in \bar{\mathcal{X}}'} p_{\mathcal{E}}(x|x') p(x') \right| \\ &\leq \left| f(x)^{(l)} - \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} p_{\mathcal{E}}(x|x') \right| + \sum_{x' \in \bar{\mathcal{X}}'} p_{\mathcal{E}}(x|x') |f(x')^{(l)} - p(x')|. \end{aligned}$$

Define

$$f(x, x')^{(l)} := \frac{|\{(k, w) | x_k(w) = x, x'_k(w) = x', 1 \leq k \leq l, 1 \leq w \leq 2^{lR}\}|}{l \cdot 2^{lR}}$$

for $x \in \bar{\mathcal{X}}, x' \in \bar{\mathcal{X}}'$. By using the relation

$$f(x)^{(l)} = \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} \frac{f(x, x')^{(l)}}{f(x')^{(l)}}, \quad (\text{A.13})$$

we obtain

$$\begin{aligned} \Delta(x)^{(l)} &:= \left| f(x)^{(l)} - \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} p_{\mathcal{E}}(x|x') \right| \\ &\leq \sum_{x' \in \bar{\mathcal{X}}'} f(x')^{(l)} \left| \frac{f(x, x')^{(l)}}{f(x')^{(l)}} - p_{\mathcal{E}}(x|x') \right|. \end{aligned} \quad (\text{A.14})$$

Applying the weak law of large numbers for each term in the sum, we have $\Delta(x)^{(l)} \rightarrow 0$ ($l \rightarrow \infty$) in probability. We also have

$$\sum_{x' \in \bar{\mathcal{X}}'} p_{\mathcal{E}}(x|x') |f(x')^{(l)} - p(x')| \leq \tau^{(l)} \cdot |\bar{\mathcal{X}}'| \quad (\text{A.15})$$

and thus

$$\lim_{l \rightarrow \infty} \sum_{x' \in \bar{\mathcal{X}}'} p_{\mathcal{E}}(x|x') |f(x')^{(l)} - p(x')| = 0. \quad (\text{A.16})$$

Therefore we obtain

$$\tau^{(l)} = \max_x |f(x)^{(l)} - p(x)| \rightarrow 0 \quad \text{in probability}. \quad (\text{A.17})$$

□

Theorem A.3 R is achievable with $p(x)$ by the channel I.

Proof. Take arbitrary $\epsilon, \delta, \eta > 0$. From Lemma A.1 and Lemma A.2, for sufficiently large l we have

$$\bar{P}_e^{(l)} < \epsilon \quad (\text{A.18})$$

and

$$\Pr\{\tau^{(l)} < \delta\} > 1 - \eta. \quad (\text{A.19})$$

Define $C_\delta^{(l)} := \{\mathcal{C}^{(l)} | \tau^{(l)} < \delta\}$. The average error probability averaged over all codebooks in $C_\delta^{(l)}$ is calculated to

$$\frac{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)}) P_e^{\mathcal{C}^{(l)}}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} = \frac{\bar{P}_e^{(l)} - \sum_{\mathcal{C}^{(l)} \notin C_\delta^{(l)}} P(\mathcal{C}^{(l)}) P_e^{\mathcal{C}^{(l)}}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} \leq \frac{\bar{P}_e^{(l)}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} < \frac{\epsilon}{1 - \eta}.$$

Thus there exists at least one codebook $\mathcal{C}^{(l)} \in C_\delta^{(l)}$ such that $P_e^{\mathcal{C}^{(l)}} < \epsilon' = \epsilon/(1 - \eta)$ and, by definition, $\tau^{(l)} < \delta$. Hence there exists a sequence of $(2^{lR}, l)$ codes for the channel I such that $P_e^{(l)} \rightarrow 0$ and $\tau^{(l)} \rightarrow 0$ when $l \rightarrow \infty$, and thus R is achievable with $p(x)$ by the channel I. \square

Appendix B. State space of a qubit

Suppose that information of two independent and uniformly random bits $x_0 x_1$ is encoded into the state of a qubit $\rho_{x_0 x_1}$. Let $\{\hat{M}_t^m\}_{t=0,1}$ be the optimal measurement for decoding x_m ($m = 0, 1$), where the mutual information $I_C(X_m : T)$ between X_m and the measurement outcome T is maximized when the measurement m is performed. We assume that for all x_0 and x_1 ,

$$P(t = x_0 | m = 0, x_0, x_1) = \text{tr}[\hat{M}_{x_0}^0 \rho_{x_0 x_1}] = \frac{1 + \alpha}{2} \quad (0 \leq \alpha \leq 1), \quad (\text{B.1})$$

$$P(t = x_1 | m = 1, x_0, x_1) = \text{tr}[\hat{M}_{x_1}^1 \rho_{x_0 x_1}] = \frac{1 + \beta}{2} \quad (0 \leq \beta \leq 1). \quad (\text{B.2})$$

In what follows, we prove that such a set of the density operators $\{\rho_{x_0 x_1}\}_{x_0 x_1}$ and POVM operators for the measurements exists if and only if $\alpha^2 + \beta^2 \leq 1$. Considering the parametrization of a qubit state using the Bloch sphere, the sufficiency is obviously verified. The necessity is proved as follows. Let $\mathbf{r}_{x_0 x_1}$ be the Bloch vector representation of $\rho_{x_0 x_1}$ and \mathbf{u}, \mathbf{v} be those of \hat{M}_0^0 and \hat{M}_1^1 , respectively. Formally, we have

$$\rho_{x_0 x_1} = \frac{1}{2}(I + \mathbf{r}_{x_0 x_1} \cdot \boldsymbol{\sigma}) \quad (\|\mathbf{r}_{x_0 x_1}\| \leq 1), \quad (\text{B.3})$$

$$\hat{M}_t^0 = \frac{1}{2}(I + (-1)^t \mathbf{u} \cdot \boldsymbol{\sigma}), \quad (\text{B.4})$$

and

$$\hat{M}_t^1 = \frac{1}{2}(I + (-1)^t \mathbf{v} \cdot \boldsymbol{\sigma}), \quad (\text{B.5})$$

where $\boldsymbol{\sigma} = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z)$. The optimality of the measurement implies that $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$. From the condition (B.1) and (B.2), we obtain

$$\begin{aligned} \mathbf{u} \cdot \mathbf{r}_{00} &= \mathbf{u} \cdot \mathbf{r}_{01} = -\mathbf{u} \cdot \mathbf{r}_{10} = -\mathbf{u} \cdot \mathbf{r}_{11} = \alpha, \\ \mathbf{v} \cdot \mathbf{r}_{00} &= -\mathbf{v} \cdot \mathbf{r}_{01} = \mathbf{v} \cdot \mathbf{r}_{10} = -\mathbf{v} \cdot \mathbf{r}_{11} = \beta. \end{aligned} \quad (\text{B.6})$$

Let $\bar{\mathbf{r}}_{x_0x_1}$ be the projection vectors of $\mathbf{r}_{x_0x_1}$ onto the two dimensional subspace spanned by \mathbf{u} and \mathbf{v} . Then we have

$$\bar{\mathbf{r}}_{00} + \bar{\mathbf{r}}_{11} = \bar{\mathbf{r}}_{01} + \bar{\mathbf{r}}_{10} = \mathbf{0} . \quad (\text{B.7})$$

and

$$\mathbf{u} \cdot (\bar{\mathbf{r}}_{00} - \bar{\mathbf{r}}_{01}) = \mathbf{v} \cdot (\bar{\mathbf{r}}_{00} - \bar{\mathbf{r}}_{10}) = 0 . \quad (\text{B.8})$$

Due to the optimality of the decoding measurement, we also have $\mathbf{u} \parallel (\bar{\mathbf{r}}_{00} + \bar{\mathbf{r}}_{01})$ and $\mathbf{v} \parallel (\bar{\mathbf{r}}_{00} + \bar{\mathbf{r}}_{10})$. Thus we obtain $\mathbf{u} \cdot \mathbf{v} = 0$. Hence

$$\alpha^2 + \beta^2 = (\mathbf{u} \cdot \bar{\mathbf{r}}_{x_0x_1})^2 + (\mathbf{v} \cdot \bar{\mathbf{r}}_{x_0x_1})^2 \leq \|\mathbf{r}_{x_0x_1}\|^2 \leq 1 . \quad (\text{B.9})$$

Appendix C. Inclusion relation of the sets of no-signalling correlations

The inclusion relation of the sets of bipartite and multipartite no-signalling correlations are given in (C.1).

$$\begin{array}{ccccccc} \mathcal{NS} = \mathcal{NSS} \supset \mathcal{IC} \supseteq \mathcal{CR} \supseteq \mathcal{Q} \supset \mathcal{C} & & & & & & (\text{C.1}) \\ (a) & (b) & (c) & (d) & (e) & & \end{array}$$

\mathcal{NS} is the set of all no-signalling correlations. \mathcal{NSS} is the set of all no-signalling correlations that satisfies no-supersignalling condition. By “satisfy” we mean that for any communication protocol using that correlation, the condition is never violated. Similarly, \mathcal{IC} and \mathcal{CR} are the sets of all no-signalling correlations that satisfy information causality and the chain rule, respectively. \mathcal{Q} and \mathcal{C} are the sets of quantum and classical correlations, respectively. \supset represents the genuine inclusion relation, and \supseteq indicates that we do not know whether the sets are equivalent or have a genuine inclusion relation. (a) is proved in Section 5. (b) is proved in [4]. (c) follows from the discussion in Section 7. (d) is obvious and (e) is proved in [1]. Recently it is proved from the observation of tripartite nonlocal correlations that at least one of (c) and (d) is a genuine inclusion relation [22, 23].

References

- [1] Bell J S 1964 *Physics* **1** 195
- [2] Clauser J F, Horne M A, Shimony A and Holt R A 1969 *Phys. Rev. Lett.* **23** 880
- [3] Popescu S and Rohrlich D 1994 *Found. Phys.* **24** 379
- [4] Pawłowski M, Paterek T, Kaszlikowski D, Scarani V, Winter A and Żukowski M 2009 *Nature* **461** 1101
- [5] Pawłowski M and Scarani V 2011 arXiv:1112.1142
- [6] Brassard G, Buhrman H, Linden N, Méthot A A, Tapp A and Unger F 2006 *Phys. Rev. Lett.* **96** 250401
- [7] Linden N, Popescu S, Short A J and Winter A 2007 *Phys. Rev. Lett.* **99** 180502
- [8] Brunner N and Skrzypczyk P 2009 *Phys. Rev. Lett.* **102** 160403
- [9] Tsirel'son B S 1980, *Lett. Math. Phys.* **4** 93

- [10] Ambainis A, Nayak A, Ta-Shma A and Vazirani U 2002 *J. ACM* **49** 496
- [11] Beigi S and Gohari A 2011 arXiv:1111.3151
- [12] Schumacher B 1995 *Phys. Rev. A* **51** 2738
- [13] Hausladen P, Jozsa R, Schumacher B, Westmoreland M and Wootters W K 1996 *Phys. Rev. A* **54** 1869
- [14] Schumacher B and Westmoreland M 1997 *Phys. Rev. A* **56** 131
- [15] Holevo A S 1998 *IEEE Trans. Inf. Theory* **44** 269
- [16] Ver Steeg G and Wehner S 2009 arXiv:0811.3771
- [17] Barrett J 2007 *Phys. Rev. A* **75** 032304
- [18] Barnum H, Barrett J, Clark L O, Leifer M, Spekkens R, Stepanik N, Wilce A and Wilke R 2010 *New J. Phys.* **12** 033024
- [19] Short A J and Wehner S 2010 *New J. Phys.* **12** 033023
- [20] Dahlsten O C O, Lercher D and Renner R 2011 arXiv:1108.4549
- [21] Al-Safi S W and Short A J 2011 *Phys. Rev. A* **84** 042323
- [22] Gallego R, Würflinger L E, Acín A and Navascués M 2011 *Phys. Rev. Lett.* **107** 210403
- [23] Yang T H, Cavalcanti D, Almeida M L, Teo C and Scarani V 2012 *New J. Phys.* **14** 013061